

P-2209

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-187753

(43)Date of publication of application : 21.07.1998

(51)Int.Cl. G06F 17/30
G06F 12/00
G06F 12/00

(21)Application number : 08-356218

(71)Applicant : NEC CORP

(22)Date of filing : 25.12.1996

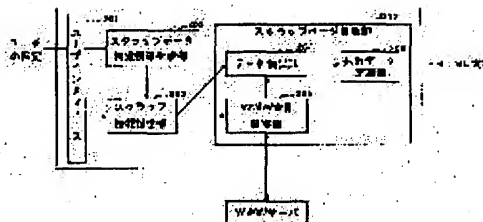
(72)Inventor : SUGIURA ATSUSHI

(54) WWW INFORMATION EXTRACTION SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide an information required by a user on a WWW with a little burden by generating and storing information for specifying the start and end position of designated data:

SOLUTION: The start and end positions of data required for a user himself are designated in a WWW document by a user interface 201, and the designated contents are specified by a scrap data specification information part 202. The pair of URL and scrap data specification information of a document, for which a user designates scrap data, is stored in a scrap information storage part 203. The latest WWW document corresponding to the stored URL is possessed by a possessing part 205 and based on the stored scrap data specification information, data required for the user are extracted out of the newly possessed WWW document by a data extracting part 204. Afterwards, the extracted data are collected into one page by a link part 206.



LEGAL STATUS

[Date of request for examination] 25.12.1996

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 2867986

[Date of registration] 25.12.1998

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

(11) 特許出願公開番号

(43)公開日 平成10年(1998)7月21日

(51)InCl ₃		識別記号		片内整理番号		F1		技術表示箇所	
G 0 6 F	17/30					G 0 6 F	15/A03	3 3 0 Z	
	12/00	5 1 5					12/00	5 1 5 A	
		5 4 7						5 4 7 H	
							15/40	3 1 0 F	
							15/A01	3 4 0 A	
						審査請求 有		請求項の数 5	F D (全 9 頁)

(21) 出版番号	特選平8-356218	(71) 出版人	000004237
(22) 出版日	平成8年(1996)12月25日		日本電気株式会社 東京都港区芝五丁目7番1号

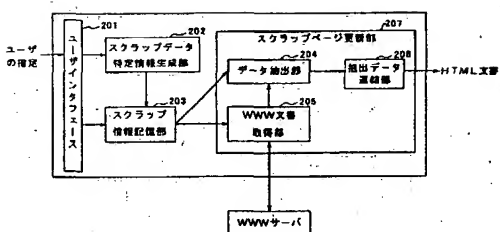
東京都港区芝五丁目7番1号 日本電気株式会社社内
(74)代理人 井理士 加藤、朝道

(54) 【発明の名称】 WWW情報抽出システム

(57) 【烟香】

【課題】日々更新されるWWW文書から必要な情報のみを効率よく少ない手間で抽出するWWW抽出システムの提供。

【解決手段】 先ずユーザがWWW文書中で必要とするデータが存在する箇所を指定し、その指定箇所を特定するための情報を生成・保存しておき、ここで生成した情報を利用して、更新されたWWW文書からユーザが必要とするデータのみを抽出する。



【特許請求の範囲】

【請求項1】 WWW (World Wide Web: ワールドワイドウェブ) サイトから情報を取得するWWW情報抽出システムにおいて、

ユーザがWWW文 について必要とするデータの存在する箇所を特定した際に、前記WWW文書における前記指定箇所を特定するための情報 (特定情報) という) を自動生成するとともに、生成された該特定情報を保存しておき、これ以降、新たに取得したWWW文書から、前記特定情報を利用して、前記ユーザが必要とするデータのみを抽出する、ことを特徴とするWWW情報抽出システム、

【請求項2】 前記指定箇所を特定するための特定情報 が、前記WWW文 において前記ユーザが指定したデータの開始および終了箇所を指示する情報からなり、該特定情報を、前記WWW文 のURL (Uniform Resource Locator: ユニフォームリソースロケータ) と対応させて保存する、ことを特徴とする請求項1記載のWWW情報抽出システム、

【請求項3】 保存されている前記特定情報に基づき、新たに取得したWWW文 中において、前記ユーザが先に指定したデータに対応するデータの開始箇所と終了箇所を判断し、前記開始箇所と前記終了箇所の間にあるデータを抽出することにより、前記ユーザが必要としていると考えられるデータを前記新たに取得したWWW文書の中から抽出する、ことを特徴とする請求項2記載のWWW情報抽出システム、

【請求項4】 保存されている全てのURLと前記特定情報の組に対して新たに取得したWWW文書の開始箇所と終了箇所の間にあるデータを抽出し、前記新たに取得したWWW文書が複数ある場合、各WWW文書から抽出したデータを一つの文書にまとめて提示する手段を備えた、ことを特徴とする請求項2記載のWWW情報抽出システム、

【請求項5】 WWW (World Wide Web: ワールドワイドウェブ) サイトから情報を取得するWWW情報抽出システムにおいて、

ユーザがWWW文書中のデータの一部分を指定することを可能とする手段と、ユーザが指定したデータの開始および終了箇所を特定するための情報を生成するデータ特定情報生成手段と、ユーザがデータを指定したWWW文書のURL (Uniform Resource Locator) および前記データ特定情報生成手段で生成された情報を記憶する記憶手段と、前記WWWサイトからWWW文書を取得するWWW文書取得手段と、

前記データ特定情報生成手段で生成され前記記憶手段に記憶されている情報を用いてユーザが先に指定したデータに対応するデータを前記WWW文 から抽出する手段と、

を備え、WWW文書の内容が更新されたとしても、ユーザが指定したデータに対応するデータを抽出する、ことを特徴とするWWW情報抽出システム、

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、WWW (World Wide Web) サイトからWWW文書を取得するシステムに関し、特にユーザが必要とする情報のみを抽出することを可能とするWWW情報抽出システムに関する。

【0002】

【従来の技術】 WWWサイトに存在する情報を取得し閲覧するためのツールとして、WWWブラウザがある。ユーザは目的のWWW文書のURL (Uniform Resource Locator: インターネットWWWアドレス) を識別するために使われる規格) を指定することにより目的の情報を閲覧することが可能である。

【0003】 また、URL指定のためのユーザの手間を簡便化する目的で、オートバイロツツールが開発されている。これを利用すると、ユーザが指定した日時や時間間隔で、ユーザが予め指定しておいた複数のURLのWWW文書をまとめて取得することが可能である。

【0004】

【発明が解決しようとする課題】 WWWブラウザを用いてWWWサイトに存在するWWW文書から必要な情報を得るためには、ユーザは、目的のWWW文書のURLを指定し、その文書がWWWブラウザ上に表示された後に、(人手による検索)、あるいは文字列検索機能を利用するといった作業を行なう必要がある。

【0005】 そして、複数のWWW文書から情報を取得する場合にも、上記の作業を繰り返すことになる。この場合には、かなりの頻度で、上記と同様の作業を繰り返す行なわなければならない、こととなる。このように、WWW文書が日々更新される場合、複数のWWW文書から情報を得るための操作が煩雑なものとなつて、ユーザの負担が増大し、この検索作業のために、所望する情報を迅速に得ることが困難となる。

【0006】 また、上記したオートバイロツツールを用いることにより、複数のWWW文書を一括したローカルマシン上へダウンロードすることが可能であるが、この場合でも、ユーザはローカルマシン上にあるファイルを自分で開き、必要な情報を得なければならぬ。

【0007】 このため、ユーザが必要な情報を閲覧するための手間は、WWWブラウザを用いる場合と、それほど変わりはない。

【0008】 したがつて、本発明は、上記した事情に鑑みてなされたものであつて、その目的は、上記従来方式の問題点を解消し、WWW上でユーザが必要とする情報

を少ない負担で得ることを可能とする、WWW情報抽出システムを提供することにある。

【0009】

【課題を解決するための手段】 前記目的を達成するため、本発明のWWW情報抽出システムは、ユーザが必要とする情報のみを複数のWWWページから自動的にスクラップし、一つの文書にまとめてユーザに提示する、ようにしたものである。

【0010】 すなわち、本発明のWWW情報抽出システムは、WWW (World Wide Web: ワールドワイドウェブ) サイトから情報を取得するシステムにおいて、ユーザがWWW文書中のデータの一部分を指定することを可能とする手段と、ユーザが指定したデータの開始および終了箇所を特定するための情報を生成するデータ特定情報生成手段と、ユーザがデータを指定したWWW文書のURL (Uniform Resource Locator) および前記データ特定情報生成手段で生成された情報を記憶する記憶手段と、前記WWWサイトからWWW文書を取得するWWW文書取得手段と、前記データ特定情報生成手段で生成され前記記憶手段に記憶されている情報を用いてWWW文書から抽出する手段と、を備え、WWW文書の内容が更新されたとしても、ユーザが指定したデータに対応するデータを抽出する、ことを特徴としたものである。

【0011】

【発明の実施の形態】 本発明の実施の形態について以下に説明する。本発明は、その好ましい実施の形態において、(a) ユーザがWWW文書中で自分が必要とするデータの開始箇所と終了箇所を指定することを可能とするユーザインタフェース (図2の201)、(b) ユーザが前記ユーザインタフェース上で指定したデータ (以下「スクラップデータ」という) の開始および終了箇所を特定するためのスクラップデータ特定情報を生成するための手段 (図2の202)、(c) ユーザがスクラップデータを指定した文書のURLおよびスクラップデータを特定情報の組 (以下「スクラップ情報」という) を記憶するためのスクラップ情報記憶部 (図2の203) を有する。

【0012】 ユーザは複数のWWWページの複数箇所に対してスクラップデータの指定を行うことを可能とする。

【0013】 システムは、ユーザが指定したそれぞれのスクラップデータに対して、スクラップデータ特定情報を生成し、スクラップ情報記憶部 (図2の203) に記憶する。

【0014】 本発明は、その好ましい実施の形態において、さらに、上記 (a) ~ (c) のに加え、(d) スクラップ情報記憶部 (図2の203) に記憶されたURLに対応する最新のWWW文 を取得する手段 (図2の205)、(e) スクラップ情報記憶部 (図2の203)

に記憶されたスクラップデータ特定情報に基づき、新たに取得したWWW文書中から、ユーザが必要とするデータの開始箇所と終了箇所を特定し、開始箇所と終了箇所の間にあるデータを抽出する手段 (図2の204)、(f) スクラップ情報記憶部 (図2の203) に記憶された全てのURLとスクラップデータ特定情報の組に対して、上記 (d) および (e) における処理を行った後に、抽出したデータを一つのページにまとめて表示する手段 (図2の206) を備えて構成されている。

【0015】 本発明は、その好ましい実施の形態において、ユーザは、最初にWWW文書において自分が必要とするデータの開始箇所と終了箇所を指定しておけば、以後はシステム間で、最新のWWW文書を取得し、新たに取得した文書からユーザが必要としていると考えられるデータのみを自動抽出する。

【0016】 このため、新たに取得したWWW文書中からユーザ自身が自分の必要とする情報を検索する必要はない。

【0017】 また、ユーザが必要とする情報の複数のWWW文書に存在する場合であっても、本発明の実施の形態に係るWWW情報抽出システムは、ユーザが必要とするデータを、各WWW文 から抽出し、抽出したデータを一つの文書にまとめてユーザに提示するため、ユーザは各WWW文書を一つ一つ開いて内容を閲覧する必要はなく、自分の必要な情報のみを一括して閲覧することが可能である。

【0018】 このため、WWWサイトから必要な情報を取得し検索するために要する作業コストを大幅に軽減することが可能である。

【0019】

【実施例】 上記した本発明の実施の形態について更に詳細に説明すべく、本発明の実施例について以下に説明する。

【0020】 図1は、本発明を一実施例のシステムの全体構成の概略をブロック図にて示したものである。図1を参照すると、本実施例に係るシステムは、データを表示するためのディスプレイ装置101と、マウスなどのポインティングデバイスおよびキーボードといった入力装置102と、スクラップデータ特定情報などを記憶するための記憶装置103と、これらディスプレイ装置101、入力装置102、及び記憶装置103を制御して処理を実行するコンピュータ104と、を備えて構成される。

【0021】 図2は、本発明の一実施例のWWW情報抽出システムの構成を機能ブロック図にて示したものである。

【0022】 図2を参照すると、WWW情報抽出システムは、ユーザがWWW文 の特定箇所を指定するためのユーザインタフェース201と、ユーザが指定したデータをWWW文書内で特定するための情報を生成するスク

ラウデータ特定情報生成部202と、ユーザがデータを指定したWWW文書のURLとスクラップデータ特定情報組を記憶するスクラップ情報記憶部203と、スクラップページ更新部207と、を備え、スクラップページ更新部207は、指定されたURLに対応するWWW文書、WWWサイトから取得するWWW文書取得部205と、スクラップデータ特定情報に基づき新たに取得したWWW文書の一部を切り出すデータ抽出部204と、抽出したデータを連結し1つの文書にまとめる抽出データ連結部206と、を備えている。

【0023】以下では、ユーザがユーザインタフェース201上で指定したデータを「スクラップデータ」、スクラップデータの開始および終了箇所をWWW文書内で特定するためにスクラップデータをスクラップデータ特定情報生成部202で生成する情報を「スクラップデータ特定情報」、ユーザがスクラップデータを指定したWWW文書のURLとスクラップデータ特定情報の組を「スクラップ情報」と呼ぶ。

【0024】スクラップデータを指定するためのインタフェース（図201）としては、ユーザが必要とするデータを含むWWW文書のURL、およびその文書中の特定データの開始および終了箇所を指定できるものであればよい。

【0025】その一例として、表示中のテキストの選択機能とするWWWブラウザで、このインタフェースとして利用可能である。

【0026】WWWブラウザをスクラップデータ特定のインタフェースとして利用した場合、図3を参照して、ユーザは、WWWブラウザにおいて、特定箇所を選択した後（仮装表示された部分を選択箇所を示す：図中ハッチングを施した領域）、選択箇所がユーザが必要とするスクラップデータであることをシステムに指示することになる。なお図3は、WWWブラウザ上のスクラップデータの選択の一例（画面表示の一例）を示した図である。

【0027】この指示の後、システムは、WWWブラウザに、WWWブラウザが現在表示しているWWW文書のURLをスクラップ情報記憶部203に記憶する。

【0028】さらに、WWWブラウザが表示中の文書の元になっているHTML（Hypertext Markup Language：ハイパーテキストマークアップランゲージ）形式の文、およびユーザがスクラップデータとして指定したデータをスクラップデータ特定情報生成部202に送る。

【0029】スクラップデータ特定情報生成部202は、該HTML文とスクラップデータを基に、スクラップデータの開始および終了箇所をWWW文、中で特定するためのスクラップデータ特定情報を生成し、スクラップ情報記憶部203に記憶する。

【0030】スクラップデータ特定情報生成部202に

おいて生成されるスクラップ情報記憶部203に記憶保持される、このスクラップデータ特定情報は、後に、新たに取得したWWW文書からユーザの必要とする情報を抽出するために、データ抽出部204で使用するものである。したがって、WWWサイトにあるWWW文書が変更された後も、その文書中に残される可能性が高い情報である必要がある。

【0031】このような情報の一例としては、HTML文書中のタグの種類や順序といった文書構成に関わるものがある。WWWサイトでは、文書構造（見出し、リストの数や順序など）はそのままで、文書が変更されることも多い。このため、スクラップデータを囲んでいるタグが、その文書内で何番目のものであるかといった情報は有用である。

【0032】また、他の例としては、スクラップデータの開始行の内容、スクラップデータの開始/終了箇所の直前/直後の行の内容がある。通常、ユーザは、WWW文書内で変更される可能性がある箇所をスクラップデータとして指定するが、WWW文書内で変更される箇所の前後の内容は変更されないことが多い。このため、スクラップデータ直前行、開始行、および直後行の内容は有用である。

【0033】本実施例では、スクラップデータ直前行、開始行、および直後行の内容をスクラップ情報記憶部203に記憶するものとする。

【0034】図5は、本実施例において、スクラップ情報記憶部203に記憶されたデータの一例を示す図である。図5を参照して、ユーザがデータを指定したWWW文書のURLに対応させて、スクラップデータ直前行、スクラップデータ開始行、スクラップデータ直後行が格納されている。

【0035】スクラップデータ直前行、開始行、および直後行の内容として、スクラップ情報記憶部203に記憶するのは、ブラウザ上に実際に表示されるデータのみとする。

【0036】すなわち、これらの行中に含まれるテキスト、画像を表示するタグ、水平線（水平線：Horizontal Rule）を表示する<HR>タグのみを記憶し、テキストを格納するタグなどは記憶しない。

【0037】HTMLのバージョンによりブラウザ上にデータを表示する効果があるタグの種類は異なるが、本実施例では、タグおよび<HR>タグのみと仮定する。

【0038】したがって、図4に示すHTML文書をWWWブラウザに表示し、図3に示すように、ユーザがスクラップデータ（図3中仮装表示部）を指定した場合、スクラップデータ開始行として、タグおよびタグを削除した「10/21 15:00更新」という文字列のみを、スクラップ情報記憶部203

に記憶する。

【0039】また、スクラップデータ直後行には、<HR>タグを記憶する。

【0040】結果として、スクラップ情報記憶部203には、スクラップデータ特定情報として、図5の第3行目に示す情報が記憶される。すなわち、スクラップデータの直前行は、「本日のトップニュース」、スクラップデータ開始行は、「10/21 15:00」、スクラップデータ直後行は、「HR」となる。

【0041】なお、図4に示すHTML文書において、<H2>タグは中見出し、
は改行、タグは斜体（イタリック）、タグは数字なしの箇条書き、は箇条書きの項目を、それぞれ指定する。

【0042】図6は、ユーザから最新WWW情報の取得要求があった場合の処理手順をPAD（problem analysis diagram：本構造チャート）にて示した図である。説明の便宜上、スクラップ情報記憶部203に記憶された1番目のURLをSurl[1]、スクラップデータ直前行の内容をSprev[1]、スクラップデータ開始行の内容をSbegin[1]、スクラップデータ直後行の内容をSnext[1]とする。

【0043】まず、WWW文書取得部205において、Surl[1]に対応する最新のWWW文書doo[1]をHTTP（Hypertext Transfer Protocol）に基づき取得する（ステップ602）。

【0044】次にデータ抽出部204において、doo[1]から切り出すデータの開始箇所B[1]をSprev[1]、およびSbegin[1]を用いて特定する（ステップ603）。

【0045】B[1]が特定できた場合（ステップ605のYes）は、さらにデータ抽出部において、doo[1]から切り出すデータの終了箇所E[1]をSnext[1]を用いて特定する（ステップ606）。

【0046】B[1]およびE[1]が共に特定できた場合には（ステップ607のYes）、B[1]とE[1]の間にあるテキストおよびそのテキストを囲む全てのタグをdoo[1]から抽出しEXT[1]に代入する（ステップ608）。

【0047】B[1]およびE[1]のいずれかが特定できなかった場合には、抽出データ無しとする。

【0048】全てのスクラップ情報として上記の処理を行った後、抽出データ連結部206において、既に抽出したデータEXT[1]（1<=i<=n）を1つのHTML文書にまとめる（ステップ609）。

【0049】図7は、データ抽出部204において、抽出するデータの開始箇所B[1]を特定するための処理手順をPADにて示したものである。

【0050】まず、doo[1]において、Sprev[1]とSbegin[1]の文字列が連続している箇所を

検索する（ステップ701）。ただし、検索は、doo[1]から、<HR>以外のタグを除去したものに對して行う。

【0051】doo[1]の先頭から文書を逐次し、最初にマッチした箇所をSbegin[1]（文字列の先頭箇所をB[1]とする（ステップ703））。

【0052】B[1]を特定できなかった場合には、Sprev[1]の文字列だけを用いて上記の検索を行い（ステップ704）、最初にマッチした箇所の次の行の先頭をB[1]とする（ステップ706）。

【0053】この検索においてもB[1]を特定できなかった場合には、さらにSbegin[1]の文字列だけを用いて上記の検索を行い（ステップ707）、最初にマッチした箇所の直前をB[1]とする（ステップ708）。

【0054】ここでもB[1]を特定できなかった場合は、最終的にB[1]を見つけれなかったものとする（ステップ709）。

【0055】データ抽出部204において抽出するデータの終了箇所E[1]は、上記の手続きで特定したB[1]からdoo[1]（1）文書の末尾までの間で検索する。

【0056】Snext[1]の文字列を用いて検索を行い、最初にマッチした箇所の直前の行の行末をE[1]とする。マッチする箇所がない場合は、E[1]は特定できなかったものとする。

システムが最新のWWW文を取得し、新たに取得した文書からユーザが必要としていると考えられるデータのみを抽出する。したがって、新たに取得したWWW文書の中からユーザ自身が自分の必要とする情報を検索する必要はない。このため、ユーザの情報抽出のための作業を特段に軽減し、利便性を特段に向上するという顕著な効果を奏する。

【0061】また、本発明によれば、ユーザが必要とする情報が複数のWWW文に存在する場合でも、システムは、ユーザが必要とするデータを各WWW文書から抽出し、抽出したデータを1つの文書にまとめてユーザに提示するため、ユーザを各WWW文書を1つ1つ開いて内容を開覧する必要はなく、自分の必要な情報のみを一括して閲覧することが可能である。したがって、本発明によれば、WWWサイトから必要な情報を取得し、検索するための作業コストを大幅に軽減することが可能である。

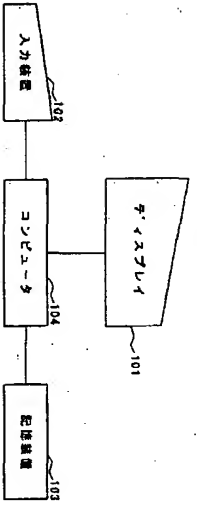
【図面の簡単な説明】

【図1】本発明の一実施例のシステム構成を示す図である。

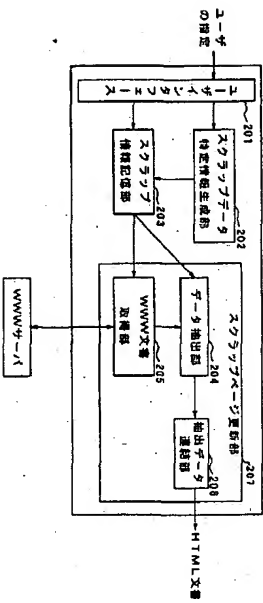
【図2】本発明の一実施例のシステムを示す図である。

【図3】本発明の一実施例を説明するための図であり、WWWブラウザ上でスクランニングデータの選択の例を示す図である。

【図1】



【図2】



HTML文書の例を示す図である。

【図5】本発明の一実施例を説明するための図であり、スクランニング情報記憶部の内容を示す図である。

【図6】本発明の一実施例を説明するための図であり、データ抽出アルゴリズムを教すPAD図である。

【図7】本発明の一実施例を説明するための図であり、データ抽出アルゴリズムを教すPAD図である。

【図8】本発明の一実施例を説明するための図であり、WWW文書の例を示す図である。

【図9】本発明の一実施例を説明するための図であり、システムにより生成されたHTML文書の例を示す図である。

【符号の説明】

- 101 表示装置
- 102 入力装置
- 103 メモリ
- 104 コンピュータ
- 105 ユーザインタフェース
- 201 スクランニングデータ特定情報生成部
- 202 スクランニング情報記憶部
- 203 データ抽出部
- 204 WWW文書取得部
- 205 抽出データ連結部

【図9】

今日のランキン

1. 夕焼け小焼け
2. お正月
3. 母さんの歌

<http://www.as.co.jp/index.html> から抽出されたデータ

10/24 10:30更新
 ・お正月
 ・夕焼け小焼け
 ・母さんの歌

【図3】

新聞ページ

今日のトップニュース

10/21 15:00更新

・日本サッカー協会

通称結果速報

昨日行われた愛媛県選の投票率は史上最低の39%。雨天にも

【図4】

通称結果速報

本日行われた愛媛県選の投票率は史上最低の39%。雨天にも

本日行われた愛媛県選の投票率は史上最低の39%。雨天にも

本日行われた愛媛県選の投票率は史上最低の39%。雨天にも

【図5】

URL	2000年10月21日	2000年10月21日	2000年10月21日
http://www.as.co.jp/index.html	<H1>	今日のトップニュース	<H1>
http://www.as.co.jp/index.html	<H2>	今日の天気	<H2>
http://www.as.co.jp/index.html	<H3>	今日の天気	<H3>

【図6】

(A) ヒットチャート

今日のランキン

1. 夕焼け小焼け
2. お正月
3. 母さんの歌

パワースタンプ

(B) 天気予報のページ

東京の天気

10/24 10:30更新

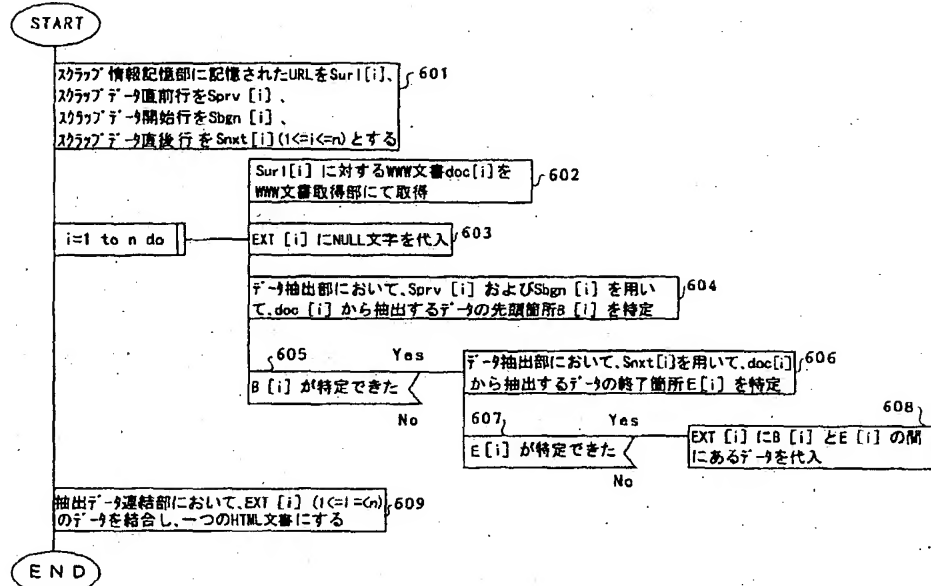
・お正月
 ・夕焼け小焼け
 ・母さんの歌

(C) 新聞ページ

今日のトップニュース

10/24 10:30更新

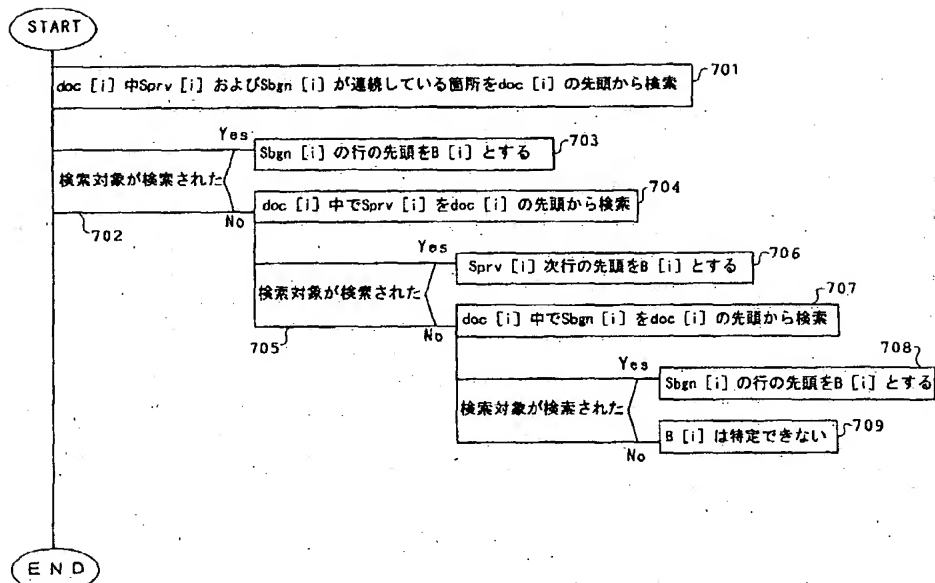
・お正月
 ・夕焼け小焼け
 ・母さんの歌



【図6】

(8)

特開平10-187753



【図7】

(9)

特開平10-187753